StorNet: Integrated Dynamic Storage and Network Resource Provisioning and Management for Automated Data Transfers

Junmin Gu[2], Dimitrios Katramatos[1], Xin Liu[1], Vijaya Natarajan[2], Arie Shoshani[2], Alex Sim[2], Dantong Yu[1], Scott Bradley[1], Shawn McKee[3]

[1]Brookhaven National Laboratory, USA
[2]Lawrence Berkeley National Laboratory, USA
[3]University of Michigan, USA

E-mail: stornet@lbl.gov

Abstract. StorNet is a joint project of Brookhaven National Laboratory (BNL) and Lawrence Berkeley National Laboratory (LBNL) to research, design, and develop an integrated end-to-end resource provisioning and management framework for high-performance data transfers. The StorNet framework leverages heterogeneous network protocols and storage types in a federated computing environment to provide the capability of predictable, efficient delivery of high-bandwidth data transfers for data intensive applications. The framework incorporates functional modules to perform such data transfers through storage and network bandwidth co-scheduling, storage and network resource provisioning, and performance monitoring, and is based on LBNL's BeStMan/SRM, BNL's TeraPaths, and ESNet's OSCARS systems.

1. Introduction

Modern data-intensive applications in the areas of high energy and nuclear physics, astrophysics, climate modelling, nanoscale materials science, genomics, etc., are expected to have an explosive growth in the amount of generated data in the next few years. Exabytes of data will have to be stored, replicated, and globally distributed to geographically distant teams of scientists for processing and analysis. Such enormous quantities of data are a major challenge to traditional data transfer tools and techniques. This is due not only to the sheer volume of data, but also to the heterogeneity of the environments at source and destination storage systems, and lack of capability to effectively utilize technology advancements in networking. The inability of traditional tools to coordinate storage and network and guarantee the end-to-end performance during a transfer leads to unpredictable transfer durations and diminished reliability, both undesirable effects, given the long times that may be required for the transfer of a large data set.

The StorNet project aims to address the deficiencies of traditional transfer tools by mating network bandwidth reservation capabilities with file transfer and storage resource management capabilities to achieve guaranteed end-to-end (disk-to-disk) data transfer performance. Recently, two major research and education networks, ESnet (run by DOE) and Internet2, were enhanced with advanced dynamic circuit switching technologies and advanced network resource reservation systems to ensure on-demand bandwidth and Quality of Service (QoS). BNL's TeraPaths system supports end-to-end network resource reservations with guaranteed quality of service by utilizing these dynamic circuit capabilities through ESnet's and Internet2's OSCARS system, and extending them within end site LAN domains. LBNL's BeStMan, provides dynamic space allocation and file management for heterogeneous storage components, including disks and tapes, and can schedule multiple concurrent file transfers. StorNet brings these three systems together in a co-scheduling framework where storage and network bandwidth can be reserved for desired time windows and dedicated to specific data

transfers, ensuring that the performance of these transfers remains predictable and within desired levels.
In the next section, we briefly describe the systems used in StorNet. Section 3 presents the architecture of the system, while section 4 focuses on the co-scheduling aspect. Section 5 describes the enhancements to the functionality of the systems used in StorNet and the StorNet interface. We summarize and talk about future directions in section 6.

2. Background

This section describes the three systems used in StorNet.

2.1. BeStMan/SRM

The Berkeley Storage Manager (BeStMan) [1] is the current LBNL's implementation of the Storage Resource Manager (SRM) standard [2-6]. SRM is an open standard that provides a common interface for middleware components to manage distributed storage systems. Several institutes have implemented and deployed SRM. The SRM specification ensures that instances of SRM interoperate with each other. BeStMan supports space reservation, dynamic space management, directory management and data transfer services. It can be adapted to support different storage systems, as well as different transfer protocols other than GSIFTP, FTP and HTTP. BeStMan also schedules multiple concurrent file transfers when possible, to efficiently use available network bandwidth.

2.2. TeraPaths

BNL's TeraPaths system [7, 8] is an end-to-end (host-to-host) network bandwidth reservation tool. TeraPaths enables end-site users/applications to reserve virtual paths with guaranteed bandwidth between pairs of end-sites. The system interacts with end-site network devices and uses differentiated services (DiffServ) and Policy-Based Routing (PBR) techniques to prioritize and dedicate bandwidth to specific network traffic within an end-site's Local Area Network (LAN) and subsequently forward this traffic into dedicated Wide Area Network (WAN) channels. These WAN channels, in the form of MPLS tunnels or dynamic circuits, are provisioned through the interfacing of TeraPaths with the OSCARS system through the latter's web services API. Invocation of TeraPaths services is also done through a web services API. OSCARS invocations are transparent to TeraPaths users.

2.3. OSCARS

The On-Demand Secure Circuits and Advance Reservation System (OSCARS) [9] is a guaranteed bandwidth provisioning system for DOE's ESnet standard IP network and advanced Science Data Network (SDN). OSCARS dynamically provisions virtual paths with guaranteed QoS. Through the collaboration between ESnet and Internet2, the system evolved into a more general Inter-Domain Controller (IDC) that provides MPLS tunnels within ESnet and guaranteed bandwidth layer 2 circuits within and between ESnet's Science Data Network (SDN) and Internet2's Dynamic Circuit Network (DCN) [10].

3. StorNet Workflow

We envision end-to-end data transfer capabilities where client applications at end-sites make requests for data residing at other end-sites, triggering a cascade of negotiation actions between storage managers and network managers. First, corresponding site BeStMan instances coordinate and determine the total volume of data to be transferred, maximum achievable bandwidth of storage systems, and feasible time frames that guarantee the data movements from client requests. Next, this information is passed to TeraPaths instances which in turn coordinate and determine a set of possible time windows during which the network can guarantee sufficient resources to satisfy BeStMan's request. Part of this process involves one more level of coordination, i.e., negotiation with the OSCARS system that manages the WAN interconnecting the end-sites. The outcome of the overall negotiation between storage and network managers is bandwidth reservations with appropriate start time so as to satisfy the transfer of the desired data volume within the time frame requested by the original client. BeStMan instances continuously monitor transfer performance, while TeraPaths monitors the health of the established end-to-end paths. Interaction between all three systems enables detection and recovery of failures by restarting failed transfers and/or re-establishing failed network paths.

The typical workflow for StorNet is depicted in Figure 1: triggered by a client's request, end-site BeStMan instances first coordinate between them to reserve storage space, and decide the parameter space (in terms of maximum bandwidth and maximum time to completion) that satisfies the request. This parameter space is then passed to TeraPaths as a request for network bandwidth reservation. TeraPaths instances coordinate between them to match the BeStMan request to LAN resource availability. Subsequently, TeraPaths generates corresponding requests for WAN bandwidth reservations and submits them to OSCARS. When multiple WAN domains are involved, OSCARS IDCs coordinate in a daisy-chain manner to establish the path interconnecting the end-sites; however, this is done transparently, i.e., TeraPaths only interacts with one IDC.

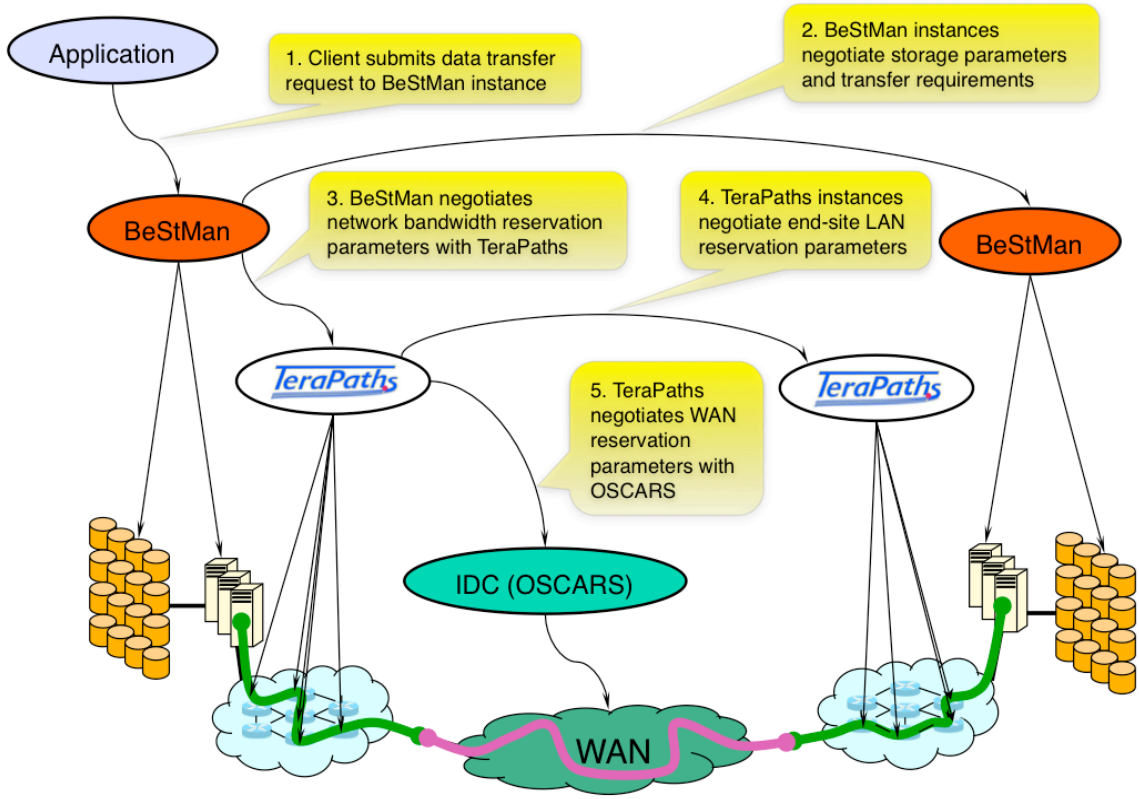

Figure 1: StorNet workflow

4. Resource Co-Scheduling

StorNet addresses a general Resource Co-Scheduling (RCS) problem: given a set of limited resources of different types and a variety of requests from data-intensive applications, determine how to optimally allocate and schedule the resources required by each application. For example, consider an application performing a time-constrained end-to-end data transfer. To reliably transfer data from source disks to destination disks over the network at known rates and meet its deadline, this application may simultaneously require a bandwidth-guaranteed network circuit and a number of dedicated CPUs and hard disks. We therefore need to jointly allocate and co-schedule all required types of resources.

For this purpose, we have developed an analytical model of resource co-scheduling, based on the concept of an end-to-end Bandwidth Availability Graph (BAG). We assume that the utilization of each resource type can be scheduled by advance reservations with specific start and end time and constant bandwidth allocation for their duration. The bandwidth allocation of such a set of reservations can be aggregated and subsequently subtracted from the maximum bandwidth availability for the overall time period to yield the BAG for the resource of interest (see Figure 2a). The maximum availability can vary with time, but typically can be considered constant, at least within known time intervals. As such, a BAG is a step function. For a storage system, for example, the maximum availability could be the total achievable transfer rate, and for a network domain the maximum achievable bandwidth.

Individual BAGs can be intersected to express the minimum availability of the initial BAGs at any given time, which provides the overall availability of resources across any number of systems (see Figure 2b). The intersection of all BAGs of source and destination storage systems and interconnecting network domains yields the end-to-end BAG.

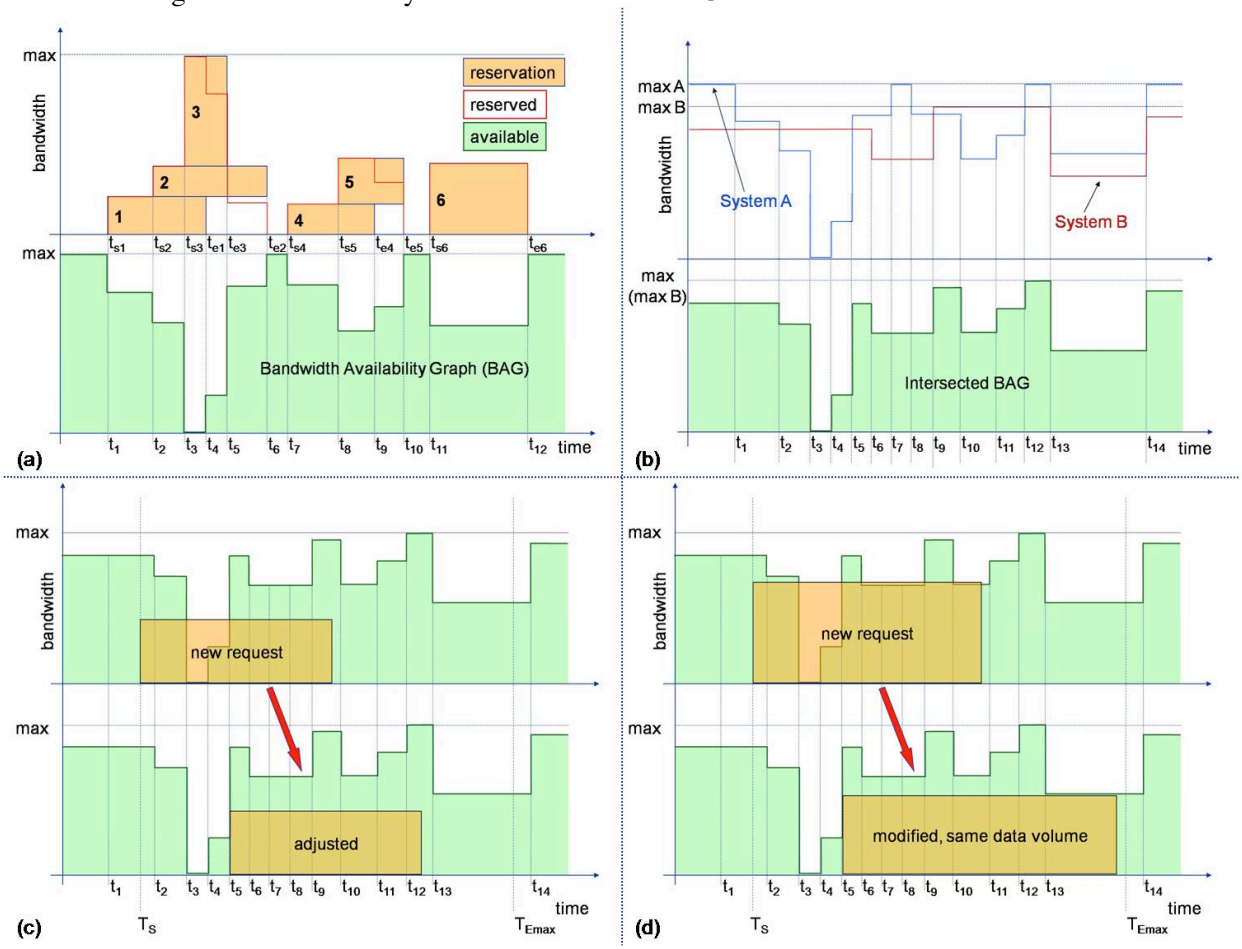

Figure 2: Bandwidth Availability Graphs

Subsequently, a new request for reserving that resource can be represented by a rectangle (see Figure 2c). If the rectangle fits into the overall BAG, then the request can be satisfied. A request may be flexible in terms of start time, duration, and/or bandwidth so that the rectangle can be modified to fit into the graph (see Figure 2d). In the latter case, the area of the rectangle represents the total volume of data to be transferred, and any modification to the start time, duration, and/or bandwidth must result in a rectangle with the same area as the initial one. The objective of fitting the request rectangle is to obtain a solution (i.e., a set of reservation parameters acceptable across all systems) that optimally satisfies the request. As optimal, we define a solution that satisfies the request according to the requestor's preferences. We have primarily considered the cases of shortest transfer duration and earliest finish time. Fitting the request rectangle can be approached as a variation of the problem of finding the largest rectangle under a histogram with n adjacent rectangles, which can be solved in $O(n)$ time using one of the known algorithms for this problem [11, 12].

5. Enhancements and Interface

StorNet approaches schedule negotiation in a top-down direction across systems, i.e., narrowing down the solution space is first performed at the BeStMan level, then at the TeraPaths level, and finally at the OSCARS level. This is done for two major reasons: firstly, because the availability of resources within each system must take into account the aspects of system-wide policies and user privileges; and secondly, because the amount of effort for figuring out a solution is reduced. The coordination of the

three systems used in the StorNet framework requires several enhancements and modifications to BeStMan and TeraPaths:

- The enhanced BeStMan can communicate with the underlying TeraPaths components, and negotiate the best solution for user requests. Through the APIs, users can specify whether they prefer earlier time solutions or shortest transfer time (i.e. higher bandwidth in a shorter time to avoid transient failures). They provide BeStMan with a desired time of completion. The BeStMan at the target site (pulling the data) communicates with the source BeStMan to find out what is its bandwidth availability. The source BeStMan returns the availability "graph" for the requested period of interest (i.e. till maximum time), in the form of a sequence of windows. The target BeStMan then finds a common schedule, and provides that to TeraPaths.
- The enhanced TeraPaths can interact with BeStMan and also supports negotiation between end-site instances through BAGs, calculation of solutions spaces by fitting requests into intersected BAGs, and negotiation with OSCARS by applying a trial-and-error approach on the set of candidate solutions obtained from the fitting process. BAG-based negotiation is not possible with the current implementation of OSCARS, however, we expect that this will be possible in the future.

The BeStMan-TeraPaths web service interface provides the functions necessary for a BeStMan server to request network bandwidth from a TeraPaths service. The goal of the API is to enable BeStMan to negotiate bandwidth with TeraPaths. The main functionalities reflected in the interface are bandwidth reservation, commitment, modification, and cancelation. The interface also includes request status checking and time-out extension. Necessary information, such as data volume, source and target resource availability, resource time frames, and other attributes, is provided to TeraPaths when requesting network bandwidth. A typical scenario is that BeStMan first attempts a temporary network bandwidth reservation. If such a reservation is possible, TeraPaths returns a request token, along with expiration time and available windows for the available resources. Once BeStMan determines that it can work with the result from TeraPaths (while determining this, BeStMan can request time-out extensions if necessary), it commits the reservation to lock the network bandwidth. Otherwise, BeStMan modifies the request parameters and resubmits the request to TeraPaths.

6. Summary and Future Work

The explosive growth of the data volumes that are collected and shared by modern data-intensive application communities underlines the need for effective and robust data transfer tools that can overcome the deficiencies of current transfer technologies. StorNet mates file and storage management capabilities with network flow prioritization and bandwidth reservation capabilities by enabling BeStMan/SRM to interface with TeraPaths and reserve on-demand network bandwidth matching the needs of data transfers. At the core of StorNet is an efficient co-scheduling scheme based on BAG intersection and request fitting algorithms that highly increases the request acceptance rate when compared with the traditional accept/reject approach in a multiple-system multiple-domain environment. Several enhancements in the BeStMan and TeraPaths systems were necessary to implement StorNet functionality, as well as BeStMan-to-TeraPaths web service-based API.

The current StorNet design targets reservations of a single transfer window with constant bandwidth per request. A future direction for the project will be to allow multiple transfer windows per request, each at a different bandwidth level, providing bigger flexibility in exploiting available resources.

7. References

[1] Berkeley Storage Manager: http://sdm.lbl.gov/bestman/

[2] Shoshani, A., Sim, A., and Gu, J.: Storage Resource Managers: Middleware Components for Grid Storage. In: Nineteenth IEEE Symposium on Mass Storage Systems, 2002

[3] Shoshani, A., Sim, A., and Gu, J.: Storage Resource Managers: Essential Components for the Grid. Chapter in book: Grid Resource Management: State of the Art and Future Trends, Edited by J. Nabrzyski, J. M. Schopf, and J. Weglarz, Kluwer Academic Publishers, 2003

[4] A. Sim, A. Shoshani, F. Donno, J. Jensen: Storage Resource Manager Interface Specification V2.2 Implementations Experience Report. GFD.154, Open Grid Forum, Aug. 2009

[5] A. Sim, A. Shoshani (Editors): The Storage Resource Manager Interface Specification Version 2.2. GFD.129, Open Grid Forum, Document in Full Recommendation, Feb. 2008
[6] Lana Abadie, Paolo Badino, Jean-Philippe Baud, Arie Shoshani, Alex Sim, et al.: Storage Resource Manager version 2.2: design, implementation, and testing experience. In: Conference for Computing in High Energy and Nuclear Physics, 2007
[7] D. Yu and D. Katramatos, TeraPaths: http://www.terapaths.org
[8] Dimitrios Katramatos, Bruce Gibbard, Dantong Yu, Shawn McKee: TeraPaths: End-to-End Network Path QoS Configuration Using Cross-Domain Reservation Negotiation. In: 3rd International Conference on Broadband Communications, Networks, and Systems (BROADNETS 2006), 2006
[9] OSCARS: https://oscars.es.net/OSCARS/docs/
[10] Internet2 Dynamic Circuit Network (DCN): http://www.internet2.edu/network/dc/
[11] Carroll Morgan: Chapter 21:The Largest Rectangle under a Histogram. Programming from Specifications, 2nd edition, Prentice Hall International (UK) Limited, October 1998
[12] ACM Collegiate Programming Contest 2003/2004: http://www.informatik.uni-ulm.de/acm/Locals/2003/html/judge.html

Acknowledgments
This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 (to LBNL) and Contract No. DE-AC02-98CH10886 (to BNL).